表層表現中の情報に基づく文章構造の自動抽出

黒橋　禎夫[†]　長尾　眞[†]

テキストや談話を理解するためには，まずその文章構造を理解する必要がある．文章構造に関する従来の多くの研究では，解析に用いられる知識の問題に重点がおかれていた．しかし，量的/質的に十分な計算機用の知識が作成されることはしばらくの間期待できない．本論文では，知識に基づく文理解という処理を行なわずに，表層表現中の種々の情報を用いることにより科学技術文の文章構造を自動的に推定する方法を示す．表層表現中の情報としては，種々の手がかり表現，同一/同義の語/句の出現，2文間の類似性，の3つのものに着目した．実験の結果これらの情報を組み合わせて利用することにより科学技術文の文章構造のかなりの部分が自動的に推定可能であることがわかった．

キーワード: 文脈処理，文章構造，結束関係，表層表現

Automatic Detection of Discourse Structure by Checking Surface Information in Sentences

Sadao Kurohashi [†] and Makoto Nagao[†]

To understand a text or dialogue, one must track the discourse structure. While work on discourse structure has mainly focused on knowledge employed in the analysis, detailed knowledge with broad coverage availability to computers is unlikely to be constructed for the present. In this paper, we propose an automatic method for detecting discourse structure by a variety of keys existing in the surface information of sentences. We have considered three types of clue information: clue expressions, occurrence of identical/synonymous words/phrases, and similarity between two sentences. Experimental results have shown that, in the case of scientific and technical texts, considerable part of the discourse structure can be estimated by incorporating the three types of clue information, without performing sentence understanding processes by giving knowledge to computers.

KeyWords: *Discourse, Coherence relation, Surface information*

1　はじめに

テキストや談話を理解するためには，**文章構造**の理解，すなわち各文が他のどの文とどのような関係 (**結束関係**) でつながっているかを知る必要がある．文章構造に関する従来の多くの研究 (Grosz and Sidner 1986; Hobbs 1979, 1985; Zadrozny and Jensen 1991, など) では，文章構造の認識に必要となる知識，またそれらの知識に基づく推論の問題に重点がおかれていた．しかしそのような知識からのアプローチには次のような問題があると考えられる．

† 京都大学工学部 電気工学第二教室, Department of Electrical Engineering, Kyoto University

- 辞書やコーパスからの知識の自動獲得，あるいは人手による知識ベース構築の現状をみれば，量的/質的に十分な計算機用の知識が作成されることはしばらくの間期待できない．
- 一方，オンラインテキストの急増にともない，文章処理の技術は非常に重要になってきている (MUC-4 1992)．そのため，現在利用可能な知識の範囲でどのような処理が可能であるかをまず明らかにする必要がある．
- 現在の自然言語処理のターゲットの中心である科学技術文では，文章構造理解の手がかりとなる情報が表層表現中に明示的に示されていることが多い．科学技術の専門的内容を伝えるためにはそのように明示的表現を用いることが必然的に必要であるといえる．

このような観点から，本論文では，表層表現中の種々の情報を用いることにより科学技術文の文章構造を自動的に推定する方法を示す．文章構造抽出のための重要な情報の一つは，多くの研究者が指摘しているように「なぜなら」,「たとえば」などの**手がかり語**である (Cohen 1984; Grosz and Sidner 1986; Reichman 1985; 小野, 浮田, ・天野 1989; 山本, 増山, ・内藤 1991, など)．しかし，それらだけで文章全体の構造を推定することは不可能であることから，我々はさらに 2 つの情報を取り出すことを考えた．そのひとつは同一/同義の語/句の出現であり，これによって**主題連鎖/焦点-主題連鎖**の関係 (Polanyi and Scha 1984) を推定することができる．もうひとつは 2 文間の類似性で，類似性の高い 2 文を見つけることによってそれらの間の**並列/対比**の関係を推定することができる．これらの 3 つの情報を組み合わせて利用することにより科学技術文の文章構造のかなりの部分が自動推定可能であることを示す．

2 文章構造のモデルと結束関係

従来，文章構造のモデルとしてはその基本単位の結束関係 (2 項関係) を再帰的に組み合わせることによる木構造 (文章構造木) が一般的に用いられてきた (Cohen 1984; Dahlgren 1988; Grosz and Sidner 1986; Halliday and Hassan 1976; Hobbs 1979, 1985; Lockman and Klappholz 1980; Mann 1984; Polanyi and Scha 1984; Reichman 1985; Zadrozny and Jensen 1991, など)．しかし，何をその基本単位とするか，また基本単位間にどのような結束関係を考えるかについては研究者ごとに独自の定義が与えられてきた．

本論文では，文章構造の自動抽出の可能性を示すことを第一義的に考え，句点で区切られた文を基本単位とするもっとも単純な文章構造モデルを採用する[1]．一方，結束関係としてどれだけのものを考えればよいかという問題は，対象とするテキストの種類に大きく依存する (Reichman 1985)．たとえば，物語文などでは過去の事象間の時系列の関係が中心となるが，科学技術文や論説文などではそのような関係はほとんどみられない．本論文では，従来の研究で扱われてきた種々の結束関係のうち，対象とする科学技術文の構造を説明するために必要なも

1 複文内の節の間にもある種の結束関係が存在すると考えられるが，その問題については本手法の発展させるかたちで別の機会に扱う．

表 1　結束関係

並列	Si と Sj が同一または同様の事象，状態などについて述べられている (例:付録 A の S4-3 と S4-6).
対比	Si と Sj が対比関係にある事象，状態などについて述べられている (例:付録 Aの S3-3 と S3-4).
主題連鎖	Si と Sj が同一の主題について述べられている (例:付録 Aの S1-13 と S1-19).
焦点-主題連鎖	Si 中の主題以外の要素 (焦点要素) が Sj において主題となっている (例:付録 Aの S1-12 と S1-13).
詳細化	Si で述べられた事象，状態，またはその要素についての詳しい内容が Sj で述べられている (例:付録 Aの S1-16 と S1-17).
理由	Si の理由が Sj で述べられている (例:付録 Aの S1-13 と S1-14).
原因-結果	Si の結果 Sj となる (例:付録 Aの S1-17 と S1-18).
変化	Si の状態が Sj のものに (通常時間経過に伴い) 変化する (例:付録 Aの S1-11 と S1-12).
例提示	Si で述べられた事象，状態の具体例の項目が Sj で提示される (例:付録 Aの S1-13 と S1-16).
例説明	Si で述べられた事象，状態の具体例の説明が Sj で行なわれる，
質問-応答	Si の質問に対して Sj で答が示される (例:付録 Aの S4-1 と S4-2).

(Si はある結束関係で接続される 2 文のうちの前の文，Sj は後ろの文を指す)

のとして表 1に示すものを考える (付録 Aに示した実験文の文章構造を図 1に示す[2]).

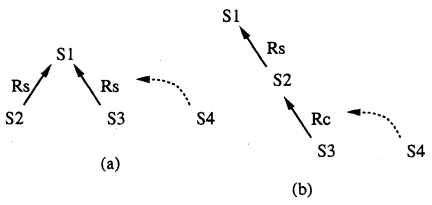
さらに，文章構造モデルに対して以下の仮定を行なうことにする.

新たな文 (入力文) は，それまでの文章構造木の右端の節点に対応する文のいずれかに接続される.

これは,「新しい主題が導入された後は，古い主題に関する詳しい説明は参照されない」ということを意味する (図 2). 計算量の点で有効であり，また科学技術文の場合直観的に妥当であると考えられたことからこの仮定を採用した[3]. なお，本論文で扱う実験テキスト (約 200 文) の各章 (9 章) についてはこの仮定のもとでそれぞれに適当な文章構造木を考えることが可能であった. 以下ではある入力文に対してそれが接続される文を**接続文**，また，接続文になり得る文章

2 図 1中，初期節点とは文章構造の初期状態として与えられるもので，実際の文には対応しない. 初期化関係はこの初期節点との間の特別の関係である. これらの詳細は 3.1 節で述べる.

3 S1 と S2 が並列/対比以外の結束関係 (Rs) を持ち，S2 と S3 が並列あるいは対比の関係 (Rc) を持つ場合，S1 と S3 が関係 Rs を持つことが推論できる. この時その次の文の S4 は，S1 と S3 だけでなく S2 との間に何らかの結束関係を持つことも考えられる. 前に示した文章構造モデルに対する仮定のもとでこの S2 への接続を許すようにするため，S3 は Rs の関係によって S1 に接続する (図 a) のではなく，単に Rc の関係で S2 に接続する (図 b) ことにする.



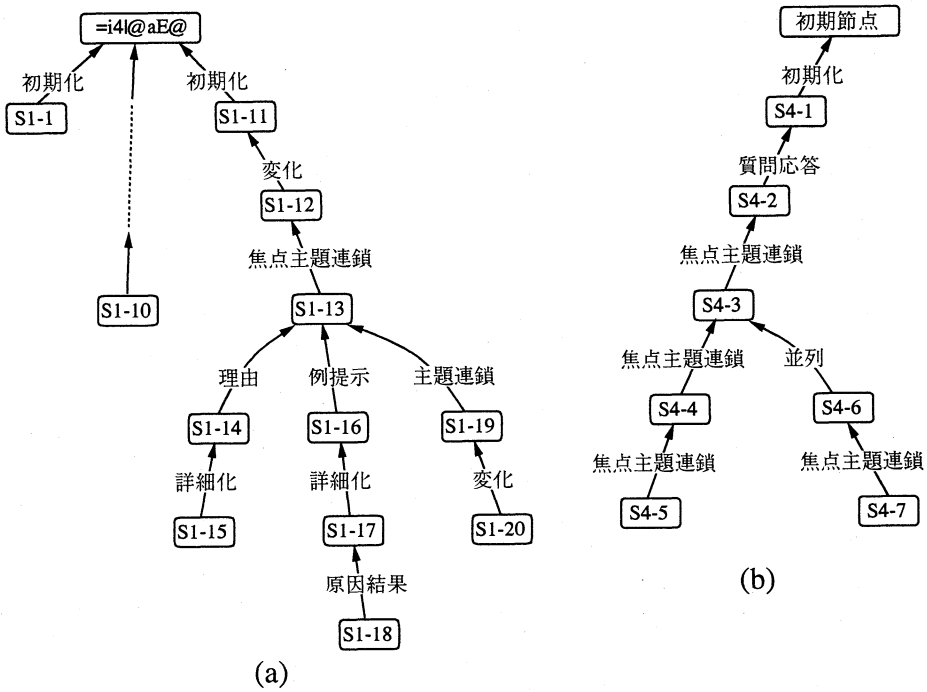


図 1 文章構造の一例

構造木の右端の節点に対応する文を**接続候補文**とよぶことにする.

3 文章構造の自動抽出

3.1 概要

前節で示したモデルに基づくと文章構造の解析は，入力文章を前から 1 文づつ順に処理し，各文 (入力文) に対して適切な接続文と適切な結束関係を決定するという問題になる．この処理を行うために，表層表現中の次の 3 つの情報に着目する.

- 種々の結束関係を示す手がかり表現
- 主題連鎖または焦点-主題連鎖関係における同一/同義の語/句の出現

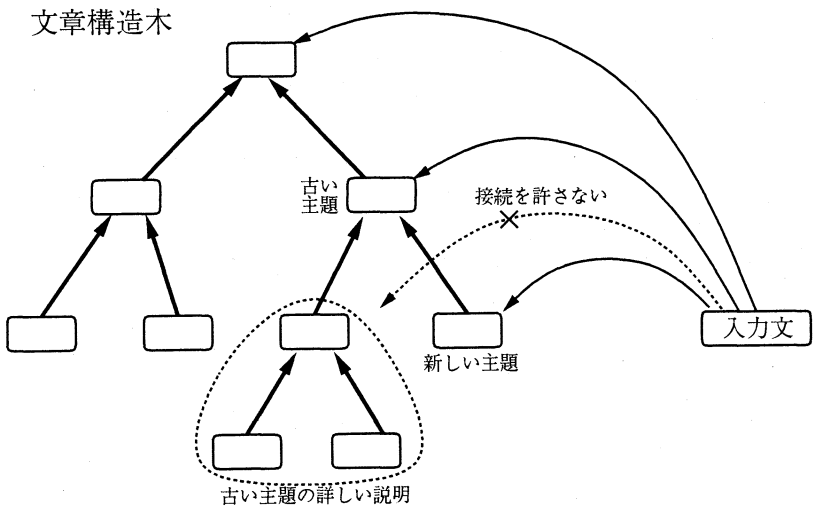


図 2　文章構造モデルに対する仮定

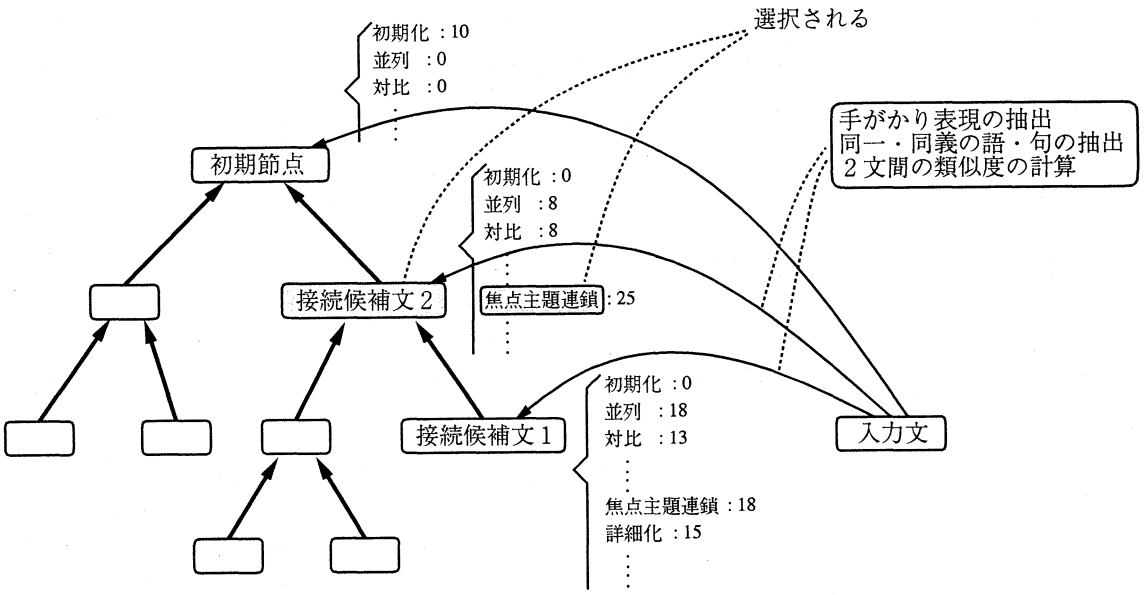


図 3　3つの表層情報による接続候補文への結束関係の得点付け

- 並列/対比関係にある2文の間の類似性

以降に示す方法によって，入力文と接続候補文に対してこれらの情報を自動的に抽出し，対応する結束関係への確信度に変換することができる．この処理によって入力文と各接続候補文との間のすべての結束関係に対する確信度を計算し，最終的に最大の確信度をもつ接続候補文と結束関係を選択する (図3).

文章構造には初期状態として**初期節点** (文には対応しない) を与え，この初期節点と入力文の間の**初期化**関係に一定の確信度を与えておく (入力文と初期節点の組については上で述べた情報の抽出処理は行なわない)．いずれの接続候補文に対してもこの値よりも大きな確信度を持つ結束関係が存在しない場合には，その入力文は初期節点に接続されるとする．これはその入力文が新しい段落の始まりの文であるような場合に対応する．

以下，各情報に対してそれらを抽出し結束関係への確信度に変換する方法を説明する．

3.2 手がかり表現の抽出

種々の結束関係を示す手がかり表現を取り出しその関係への確信度を得るために，ヒューリスティック・ルールを用意した．ルールは以下のものからなる．

- **ルールの適用条件**：
  - **ルールの適用範囲** (どれだけ離れた接続候補文までルールを適用するか)
  - **接続候補文とその接続文との結束関係**[4]
  - **接続候補文の依存構造のパターン**
  - **入力文の依存構造のパターン**
- **対応する結束関係と確信度**

接続候補文と入力文のパターンはそれぞれの文の依存構造解析結果に対して適用される (黒橋・長尾 1994)．この処理は依存構造木に対する柔軟なパターン照合機能を用いて実現した．そこでは，依存構造木とその構成要素である文節 (単語の並び) に対するパターンが，正規表現，論理和，論理積，否定などによって指定できる (Murata and Nagao 1993)．ルールは各接続候補文と入力文の組に対して適用され，条件部が満たされれば対応する結束関係に指定された確信度の得点が与えられる (複数のルールがマッチした場合確信度の得点は加算されていく).

ルールの一例を表2に示す (すべてのルールを付録Bに示す)．たとえば，ルール a(表2) は入力文が「なぜなら」で始まる場合，その入力文と直前の接続候補文の間の理由関係に得点を与える (ルールの適用範囲が1であるので，直前の接続候補文との間の関係に対してのみ得点が与えられる)．ルール b(表2) の場合は，条件部で同一の語の出現を指定しており，直前の接続候補文だけでなく他の接続候補文に対しても適用される．ルール c(表2) では，条件として「接続

4 ある入力文に対する処理を行なう時点では，それより前の部分の文章構造はすでに決定されている．そこで，接続候補文がどのような結束関係でそれ以前の文 (接続候補文の接続文) と接続されているかということをルールの適応条件とすることができる．

表 2　手がかり表現に対するルール

ルール a

適用範囲：1
接続候補文の結束関係：＊
接続候補文のパターン：＊
入力文のパターン：
結束関係：理由
確信度：20

ルール b

適用範囲：＊
接続候補文の結束関係：＊
接続候補文のパターン：＊ → X＊ → ＊
入力文のパターン：＊ → X の → 例＊ → ＊
結束関係：例提示
確信度：30

ルール c

適用範囲：1
接続候補文の結束関係：例提示
接続候補文のパターン：＊
入力文のパターン：＊
結束関係：詳細化
確信度：25

ルール d

適用範囲：1
接続候補文の結束関係：＊
接続候補文のパターン：＊ → 用言：現在形
入力文のパターン：＊ → 用言：過去形
結束関係：＊
確信度：−15

「A → B」は A が B に係る依存関係を示す．「＊」はワイルドカードを示す．

候補文とその接続文との結束関係」を指定している．このルールは，「例提示関係によって具体例が導入されれば，次にその例の説明が続く場合がある」ことを表現している．ルール d(表 2)は時制の変化を手がかりとして文章構造の区切れを検出するためのルールで，連続する 2 文の時制が現在から過去に移行する場合，その間の全ての結束関係の確信度を指定された値だけ減少させる．このペナルティの値によって入力文が直前の文以外の接続候補文へ接続されることが優先される．

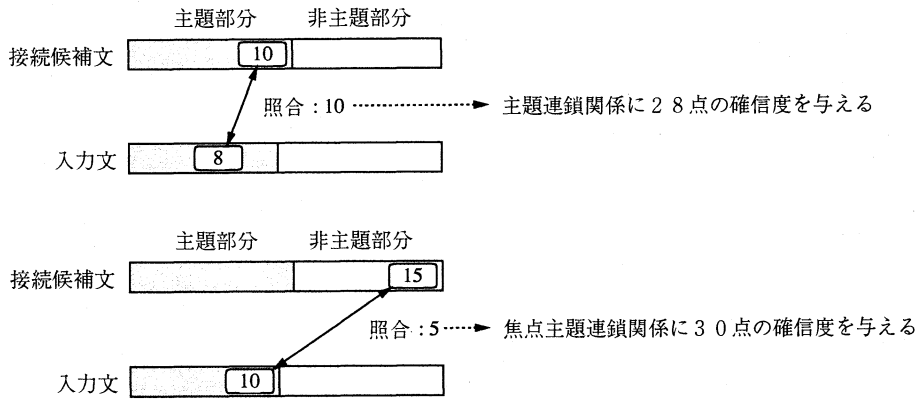


図 4 主題連鎖関係，焦点-主題連鎖関係の確信度

3.3 語連鎖の抽出

一般に文は主題を示す部分 (主題部分) とそれ以外の部分 (非主題部分) に分けることができる．2つの文が同じ主題について述べられている場合，それら2文は主題連鎖関係にあるとする．この関係は2つの文の主題部分に同一/同義の語/句 (以下これを**語連鎖**とよぶ) が現れていることで発見できる．一方，ある文の主題以外の要素がその後の文の主題となるような結束関係を焦点-主題連鎖関係とよぶことにする (この時，後の文で主題となる要素は前の文において焦点要素であると考えられるため)．この関係は2文間の主題部分から非主題部分への語連鎖を調べることで発見できる．

しかし，多くの場合一つの入力文に対して各接続候補文との種々の関係を支持する複数の情報が存在する．そのため，単に語連鎖を発見するだけでなく，その連鎖の強さに応じて主題連鎖関係あるいは焦点-主題連鎖関係に確信度を与えることが必要となる．そこで，主題部分，非主題部分の各語に対して，文中での重要度に応じた得点を与え，また同一/同義の語/句の照合に対してもその一致度に応じた得点を定義した．その上で，連鎖する2語/句のそれぞれの文での重要度の得点とその一致度の得点の総和を語連鎖の得点とし，これを主題連鎖関係あるいは焦点-主題連鎖関係に確信度として与えるということを行った (図 4)．

これらの処理は，依存構造に対するパターンと得点からなるルールを適用するという形で実現した (ルールの一例を表 3に示す)．ルール a,b(表 3) は主題を示す助詞「は」を伴う語とその修飾語/句に，ルール c,d(表 3) は条件節内の語に，主題部分としての得点を与える．ここでは，主題部分の中でもっとも重要であると考えられる，助詞「は」を直接ともなう語に最大の得点を与えている．ルール e(表 3) は「～A がある」という文の A に対して非主題部分の要素とし

表 3　主題部分，非主題部分，語/句の一致に対するルール

主題部分

ルール a

パターン：[*は]

得点：10

ルール b

パターン：[*] → * → *は

得点：8

ルール c

パターン：[用言：条件形]

得点：5

ルール d

パターン：[*] → * → 用言：条件形

得点：5

非主題部分

ルール e

パターン：[*が] → ある

得点：11

語 / 句の一致

ルール f

パターン：X * ⟷ x *

得点：5

ルール g

パターン：X * → Y *　⟷　x * → y *

得点：8

ルール h

パターン：XY* ⟷ x * → y *

得点：6

主題部分，非主題部分のルールについては□で囲まれた文節に得点を与える．
語 / 句の一致のルールでは，X と x，Y と y はそれぞれ同一の語，または分類語彙表 (国立国語研究所) のコードが 6 桁一致する語を表わす．

て高い得点を与える．このような文体では A が重要な新情報であり，この語が以降の文で主題として取り上げられる (焦点-主題連鎖関係が存在する) 場合が多い．ルール f,g,h(表 3) は語/句の一致に対して得点を与えるルールである．ここでは「X の Y」のような句の一致 (ルール g) に対しては単なる語の一致 (ルール f) よりも高い得点を与える．

語連鎖の検出における最大の問題は，著者が，単に同一の語/句を繰り返すのではなく，微妙に異なった表現によってそのような連鎖を示す傾向にあるという問題である．シソーラスの利用，ルール h(表 3) などによってそのような表現の一部は検出可能であるが，検出できない微妙な表現は多数存在する．それらの取扱いについては本論文の範囲外とした．

3.4 2 文間の類似度の計算

並列/対比の結束関係にある 2 文の間にはある種の類似性が認められる．しかしそれらは文全体についての類似性であるため，これまでに示したような文中の比較的狭い部分を調べるルールを適用することでは検出できない．そこで，1 文内の並列構造の範囲を調べるために開発したダイナミックプログラミングによる類似性検出の方法 (Kurohashi and Nagao 1992) を拡張するということを行った．

この方法では任意の長さの文節列間の類似度を計算することが可能である．そこでは，まず文節間の類似度を，品詞の一致，語の一致，シソーラス辞書中での語の近さなどによって計算し，その上でそれらの文節間の類似度を組み合わせることによって文節列間の類似度を計算する．この手法は並列構造の構成要素を決定するために 1 文内の句/節間の類似度を計算するものとして開発した．これを 2 文間 (入力文と接続候補文) の類似度を計算するように拡張することは，その 2 文を連結しそれらが仮想的に並列構造を構成すると見なすことによって簡単に実現することができる．このことにより，その仮想的並列構造の構成要素である 2 文間の類似度をまったく同じ枠組みで計算することができるからである．最終的には，この方法によってえられた類似度を 2 文の長さの和によって正規化した値を，それらの間の並列関係と対比関係の確信度に加算する．

4 実験と考察

4.1 実験方法と結果

実験には科学雑誌サイエンスのテキスト,「科学技術のためのコンピューター」(Vol.17,No.12) を用いた (付録 Aにその一部を示す)．実験文はあらかじめ章ごとに分割し (全 9 章，1 章は平均 24 文)，文章構造の推定は章単位で行なった．

実験は以下の手順で行なった．まずはじめの 3 章に対して．主題部分/非主題部分の重要度のルール，語/句の照合ルール，手がかり表現のルールを作成し，できるだけ正しい文章構造が求まるようにそれらの得点を人手で調整した[5]．2 文間の類似度の計算については，単に並列構造検出のシステムを用いただけで，得点付けの調整/変更は行なわなかった．次に，残りの 6 章に対して手がかり表現のルールだけを追加し，その新しいルールセットによって残り 6 章に対す

5 これまでに表，付録等で示したルールの得点はこの調整後のものである．また，あらかじめ初期節点との初期化関係に与えておく確信度は 10 点とした．

表 4　実験結果

結束関係	学習サンプル (はじめの 3 章)		テストサンプル (残りの 6 章)	
	正解	誤り	正解	誤り
初期化	7	1	6	2
並列	10	1	15	2
対比	6	1	2	2
主題連鎖	13	1	21	5
焦点-主題連鎖	10	4	37	14
詳細化	9	1	9	1
理由	3	0	1	0
原因-結果	2	0	6	0
変化	3	0	0	0
例提示	1	0	0	0
例説明	3	0	2	0
質問-応答	1	0	1	0
合計	68 (正解率 88%)	9	100 (正解率 79%)	26

(各章のはじめの文は必ず初期化関係となるのでこの表からは除外した)

る解析実験を行なった．これらの実験結果を表 4に示す．ここでは，各入力文に対して正しい接続文と正しい結束関係，また結束関係が主題連鎖または焦点-主題連鎖の場合はその正しい語連鎖が求まった場合を正解とした．結束関係ごとの解析結果の集計では，解析失敗のものについてはその正しい結束関係の欄に分類した．なお，残りの 6 章に対するルールを追加した新しいルールセットでもとの 3 章を解析したところ，解析結果は全く同じであった．

これらの結果から，科学技術文の場合，表層表現中の情報をうまく取り出すことができればその文章構造のかなりの部分が自動的に推定可能であることがわかる．手がかり表現についての汎用性のあるルールセットを用意するためには，かなりの規模のテキストを対象としたルール作成作業が必要であると思われる．しかし，手がかり表現に関するルールの多くは排他的なものとして記述することができるので，本実験においてそうであったように，新しく追加したルールがもとのルールと競合して副作用を起こすということは非常に少ないと予想される．

4.2　解析例と考察

まず，文章構造推定の経緯の具体例を示す．付録の S1-11 から S1-20 までの文章は以下に示す情報によって図 1-a に示す構造に変換された．

S1-11 –初期化→ 初期節点　— S1-10 との関係に対するペナルティ (付録 Bルール 1)．他の接続候補文との間に大きな確信度をもつ結束関係がないため，初期節点に接続される．

S1-12 –変化→ S1-11　— 過去から現在への時制の変化と，手がかり語「しかし」(ルール 40)．

S1-13 –焦点-主題連鎖→ S1-12　— 「合成による分析」と「合成法」の連鎖．

S1-14 –理由→ S1-13 — 手がかり表現「からである」(ルール 34).
S1-15 –詳細化→ S1-14 — 手がかり表現「わけである」(ルール 20).
S1-16 –例提示→ S1-13 — 手がかり表現「～の例として」(ルール 41).
S1-17 –詳細化→ S1-16 — 直前の例提示関係 (ルール 26).
S1-18 –原因-結果→ S1-17 — 手がかり表現「その結果は」(ルール 36).
S1-19 –主題連鎖→ S1-13 — 「合成法」の連鎖.
S1-20 –変化→ S1-19 — S1-12 と同様.

また付録の S4-1 から S4-7 までの文章は以下の手順で図 1-b に示す構造に変換された.

S4-2 –質問-応答→ S4-1 — 手がかり表現「～か」(ルール 43).
S4-3 –焦点-主題連鎖→ S4-2 — 「連星」の連鎖.
S4-4 –焦点-主題連鎖→ S4-3 — 「温めることがある」と「この過程」の連鎖 (ルール 19).
S4-5 –焦点-主題連鎖→ S4-4 — 「核融合」の連鎖.
S4-6 –並列→ S4-3 — s4-3 との類似度と手がかり表現「また」(ルール 5).
S4-7 –並列→ S4-6 — 類似度による得点が高いため誤って S4-6 との並列関係が推定された.

正解は S4-6 との焦点-主題関係である.

接続詞「しかし」は, S1-12, S1-20 のように変化関係を示す手がかり語であるだけでなく, S3-4 のように対比関係を示す場合もある. この区別は, この手がかり語と他の情報を組み合わせて調べることによって可能となる. すなわち, S1-12, S1-20 の場合は過去から現在への時制の変化を見ることによって変化関係を推定することができる. 一方, S3-3 と S3-4 の間には高い類似度 (得点 23) があるため, これと手がかり語「しかし」によって対比関係が推定できる (これに対して, S1-11 と S1-12 の類似度は 0, S1-19 と S1-20 の類似度は 3 である).

連続する 2 文間の結束関係だけでなく, 離れた 2 文間の関係も, 種々の情報によって正しく推定することが可能である. 例えば, S1-16, S1-12 間の例提示関係は表層表現中の手がかり語によって, S1-19, S1-13 間の主題連鎖関係は語連鎖によって, また S4-6 と S4-3 の並列関係は 2 文間の類似度によってそれぞれ正しく推定されている.

S4-7 については正しい結束関係が推定できなかった, ここでは S4-6 の「温度が上昇する」ことによって生じる「熱」が S4-7 で主題となっている. すなわち, 推論を介した焦点-主題連鎖関係が存在している. このように結束関係の推定に推論を必要とするような問題は本論文では対象としなかった.

このような問題を含めて, 実験テキストの解析誤りの原因は次のように分類できる. なお, 学習サンプル, テストサンプルともに解析誤りの原因の種類は同じものであった.

(1) 語連鎖の抽出について

3.3 節でも述べたように, 主題連鎖/焦点-主題連鎖関係を示す語連鎖にはシソーラスやある種のパターン (表 3のルール h など) を用いるだけでは検出不可能なものがある. 上記

のS4-7はその一例である．また逆に，主題連鎖/焦点-主題連鎖関係を示すものではない同一語句の出現に対して得点が与えられ，それが他の正しい関係の推定の妨げとなる場合がある．

(2) 2文間の類似度の計算について

たとえば，

「銀河の中には、互いに近接していて、橋がかかっているように見えるものがある。しかし大多数の銀河は、ほぼ対称的な渦巻き形、あるいは単純な円形または楕円形をしている。」

という対比関係の2文では，同一/同義の単語も少なく構造(品詞の並び)的にも似ていないため我々の方法では高い類似度が与えられない．このような場合並列/対比関係が正しく推定されないということになる．また逆に，並列/対比関係にない2文でも同一の語を多数含んでいるような場合高い類似度が与えられ，誤ってそれら2文間の並列/対比関係が推定されてしまう場合がある．

(3) 詳細化関係について

詳細化関係ではたとえば次の例のように手がかり表現といえるものが存在しない場合がある．

「こうした終局的な型は、初期条件に左右される。遭遇時の速度や傾きの角度を少し変えるだけで、複雑な3体の動きは大幅に変わる。」

このような2文間の詳細化関係は本手法では正しく推定できない．

表4に示すとおり解析誤りの大部分は並列/対比関係，主題連鎖/焦点-主題連鎖関係についての誤りで，これらの多くは上の1,2の原因が組み合わさって生じたものである．たとえば，微妙な表現の語連鎖が抽出できない場合に，別の不適当な文との間の高い類似度によって並列/対比関係が推定されてしまったというような場合である．

なお，本論文では同一著者のテキストを学習サンプル，テストサンプルとして実験を行なった．種々のテキストに対して本手法の有効性を検証することは今後の課題である．しかし，科学技術文の本質的目的がその内容を正確に伝えることであるということを考えれば，著者・テキストによって文体に差があるとしても，それは本手法の精度に大きな影響を与えるほどのものではないと考えられる．

5 結論

本論文では，手がかり表現，語連鎖，文間の類似性，という表層表現中の3つ情報に基づいて文章構造を自動的に推定する手法を示した．科学技術文の場合，これら3つの表層的情報を組み合わせて利用することにより，知識に基づく文理解という処理を行なわなくても，文章構造のかなりの部分が推定できることがわかった．微妙な表現による語連鎖の扱い，手がかり表

現についての汎用性のあるルールセットの作成などの問題を克服すれば，大規模なテキストに対して文単位でなく文章単位の処理を実現することが可能となるだろう．

参考文献

Cohen, R. (1984). "A Computational Theory of the Function of Clue Words in Argument Understanding." In *Proceedings of the 10th COLING*, pp. 251–258.

Dahlgren, K. (1988). *Naive Semantics for Natural Language Understanding.* Kluwer Academic Publishers.

Grosz, B. J. and Sidner, C. L. (1986). "Attention, Intentions, and the Structures of Discourse." *Computational Linguistics*, **12** (3).

Halliday, M. A. K. and Hassan, R. (1976). *Cohesion in English.* Longman.

Hobbs, J. R. (1979). "Coherence and Co-Reference." *Cognitive Science*, **3**, 67–90.

Hobbs, J. R. (1985). "On the Coherence and Structure of Discourse." Tech. rep..

黒橋禎夫・長尾眞 (1994). "並列構造の検出に基づく長い日本語文の構文解析." 言語処理学会, **1** (1).

Kurohashi, S. and Nagao, M. (1992). "Dynamic Programming Method for Analyzing Conjunctive Structures in Japanese." In *Proceedings of the 14th COLING*, Vol. 1, pp. 170–176.

Lockman, A. and Klappholz, A. D. (1980). "Toward a Procedural Model of Contextual Reference Resolution." *Discourse Processes*, **3**, 25–71.

Mann, W. C. (1984). "Discourse Structures for Text Generation." In *Proceedings of the 10th COLING*, pp. 367–375.

MUC-4 (1992). *Proceedings of Fourth Message Understanding Conference.*

Murata, M. and Nagao, M. (1993). "Determination of referential property and number of nouns in Japanese sentences for machine translation into English." In *Proceedings of TMI '93*, pp. 218–225.

小野顕司, 浮田輝彦, 天野真家 (1989). "文脈構造の分析." 自然言語処理研究会, 情報処理学会.

Polanyi, L. and Scha, R. (1984). "A syntactic Approach to Discourse Semantics." In *Proceedings of the 10th COLING*, pp. 413–419.

Reichman, R. (1985). *Getting Computers to Talk Like You and Me.* The MIT Press.

山本和英, 増山繁, 内藤昭三 (1991). "手がかり語を用いた日本語文章の段落分けに関する実証的考察." 自然言語処理研究会, 情報処理学会.

Zadrozny, W. and Jensen, K. (1991). "Semantics of Paragraphs." *Computational Linguistics*, **17** (2).

付録

A　実験テキスト

「科学技術のためのコンピューター」(サイエンス，Vol.17,No.12)
(以下，Si-j の i は章番号，j は文番号を示す)

S1-1: コンピューターによる高速計算は科学研究の方法を、劇的に変えつつある。

⋮

S1-10: たとえば、重力に関するニュートンの法則はすでによく理解されており、太陽系の正確な計算モデルが存在することから、計算機実験によって、火星がなければ地球の軌道がどう変わるか、といった問題を解くこともできる。
S1-11: これまでの理論家は、近似的な結果を計算して、現実の世界と比較できるようにするために、極端な単純化を強いられることが多かった。
S1-12: しかし科学者が利用できる計算機の能力が増大した１つの結果として、研究方法は従来の近似法から“合成による分析”へと移行しつつある。
S1-13: 合成法は、「あるシステムの各部分の間の相互作用の基本過程はわかっているが、当のシステムの細かな構成はわからない」という場合に使われる。
S1-14: それにより、未知の構成を合成によって決定したり、可能な構成を考えて、その結果を試してみることも可能だからである。
S1-15: そうした結果を実験から得られる細かなデータとつき合わせてみれば、観察の結果を最もよく説明できる構成を選ぶことができるわけである。
S1-16: １９世紀以来の合成法の有名な例として、天王星の軌道に見られる不可解な摂動を理解しようとした試みをあげることができる。
S1-17: 研究者たちは太陽系に仮想の惑星を加え、満足のいく摂動が得られるまで、その軌道のパラメーターを変化させていった。
S1-18: その結果は、予想された位置の近くでの海王星の発見という成果に直接結びついたのである。
S1-19: この合成法が適用できるのは、過去には、比較的単純な場合に限られていた。
S1-20: しかし、高速計算機の登場により、合成法は、伝統的な近似解析に次ぐ地位を確実に占めようとしている。

⋮

S3-3: ２つの恒星あるいは惑星間に働く重力の相互作用は、紙と鉛筆で容易に計算することができる。

S3-4: しかし、3個の物体間の相互作用（すなわち3体問題）となると、その運動方程式は手におえないものになってしまう。

⋮

S4-1: 天文学者はこの種の衝突になぜ興味をもつのだろうか。
S4-2: その答えは、“熱”を発生させるのに連星が演じている役割にある。
S4-3: 連星と単星が衝突する際、連星は縮んで小さくなり、単星にエネルギーを与え、その周囲の星の集団を温めることがある。
S4-4: この過程は、原子核が衝突して融合し、より重い原子核になる際、エネルギーを放出する核融合とよく似ている。
S4-5: 核融合は、太陽を含む恒星を光らせるメカニズムである。
S4-6: また、遭遇によって連星の軌道が縮小し、そのために高密度の星団の中核の温度が上昇することも考えられる。
S4-7: この熱は、星が絶えず沸騰している星団の表面における熱損失と釣り合うことのできるものである。

B 手がかり表現のルール

A欄：Oははじめの3章に対して作成したルール，Nは残りの6章に対して作成したルール．
適用範囲：*は制限なし．
パターン：{a|b}は「aまたはb」を示す．

No.	A	結束関係	確信度	適用範囲: 〈接続可能文の結束関係・パターン〉〈入力文のパターン〉
1	O	*	−15	1: 〈~(用言:現在形)〉〈~(用言:過去形)〉
2	O	*	−20	1: 〈~〉〈~(副助詞「は」なし)~だ(判定詞)〉
3	O	*	−20	1: 〈~〉〈~がある〉
4	O	並列	10	*: 〈~〉〈~さらに~〉
5	O	並列	15	*: 〈~〉〈~また~〉
6	O	並列	40	*: 〈[並列関係]〉〈~さらに~〉
7	N	並列	5	1: 〈~〉〈そして~〉
8	N	並列	15	1: 〈~〉〈しかも~〉
9	N	並列	30	*: 〈第~〉〈第~〉

No.	A	結束関係	確信度	適用範囲: 〈接続可能文の結束関係・パターン〉〈入力文のパターン〉
10	N	並列	40	*: 〈[並列関係]〉〈～最後に～〉
11	O	対比	15	*: 〈～〉〈～一方～〉
12	O	対比	15	1: 〈～〉〈～しかし～〉
13	O	対比	15	1: 〈～〉〈～その代わり～〉
14	O	対比	30	1: 〈[対比関係]〉〈～さらに～〉
15	N	対比	15	*: 〈～〉〈～対照的に～〉
16	N	対比	30	1: 〈～〉〈むしろ～〉
17	N	対比	40	*: 〈[対比関係]〉〈～最後に～〉
18	O	焦点-主題連鎖	30	1: 〈～(動詞){ ことがある | ことができる |(文末)}〉〈(名詞修飾形態指示詞){ 方法 | 成果 }～〉
19	N	焦点-主題連鎖	30	1: 〈～(動詞){ ことがある | ことができる |(文末)}〉〈(名詞修飾形態指示詞){ 技法 | 段階 | 過程 | サービス | 扱い }～〉
20	O	詳細化	20	1: 〈～〉〈～わけだ〉
21	O	詳細化	20	1: 〈～(数詞)～分類{ できる | される }〉〈～〉
22	O	詳細化	30	1: 〈～〉〈{ すなわち | つまり | いずれにせよ }～〉
23	O	詳細化	30	1: 〈～〉〈この場合～〉
24	O	詳細化	30	1: 〈～〉〈～とする〉
25	O	詳細化	30	1: 〈～〉〈実際～〉
26	O	詳細化	40	1: 〈[例提示関係]〉〈～〉
27	N	詳細化	20	1: 〈～〉〈まず～〉
28	N	詳細化	20	1: 〈～〉〈第一～〉
29	N	詳細化	20	1: 〈～〉〈最初の～〉
30	N	詳細化	30	1: 〈～〉〈事実～〉
31	N	詳細化	30	1: 〈～〉〈ここで～{ は | が }～だ〉
32	N	詳細化	30	1: 〈～〉〈ここで～(用言:条件形)～〉
33	O	理由	30	1: 〈～〉〈なぜなら～〉
34	O	理由	30	1: 〈～〉〈～からだ〉
35	N	理由	30	1: 〈～〉〈～{ に | にも } よる〉
36	O	原因-結果	30	1: 〈～〉〈その{ 結果 | 結果は }～〉
37	N	原因-結果	30	1: 〈～〉〈{ したがって | このため | そのため | こうすることによって | そうすることによって }～〉
38	N	原因-結果	30	1: 〈～〉〈{ こう | そう } して～ことにより～〉

No.	A	結束関係	確信度	適用範囲: 〈接続可能文の結束関係・パターン〉〈入力文のパターン〉
39	N	原因-結果	30	1: 〈~〉〈{こう | そう}して~(用言:条件形)~〉
40	O	変化	30	1: 〈~{いた(接尾辞)|(形容詞過去形)}〉〈{しかし | ところが}~〉
41	O	例提示	30	1: 〈~X~〉〈~Xの例として~{ある | あげる}〉
42	O	例説明	30	1: 〈~〉〈たとえば~〉
43	O	質問-応答	30	1: 〈~か〉〈~〉

略歴

黒橋 禎夫: 1989年京都大学工学部電気工学第二学科卒業. 1994年同大学院博士課程修了. 同年, 京都大学工学部助手, 現在に至る. 自然言語処理, 知識情報処理の研究に従事. 1994年4月より1年間Pennsylvania大学客員研究員.

長尾 眞: 1959年京都大学工学部電子工学科卒業. 工学博士. 京都大学工学部助手, 助教授を経て, 1973年より京都大学工学部教授. 1976年より国立民族学博物館教授を兼任. 京都大学大型計算機センター長 (1986.4 – 1990.3), 日本認知科学会会長 (1989.1 – 1990.12), パターン認識国際学会副会長 (1982 – 1984), 日本機械翻訳協会初代会長 (1991.3 –), 機械翻訳国際連盟初代会長 (1991.7 – 1993.7). 電子情報通信学会副会長 (1993.5).

計算機にどこまで人間的なことをやらせられるかに興味を持ち, この分野に入った. パターン認識, 画像処理, 機械翻訳等の分野を並行して研究. 機械翻訳の国家プロジェクトを率いて, 本格的な日英, 英日翻訳システムを完成した. またアナロジーの概念に基づく翻訳 (用例を用いた翻訳) を提唱. 今日その重要性が世界的に認識されるようになって来ている.

(平成4年4月4日 受付)

(平成4年5月27日 採録)